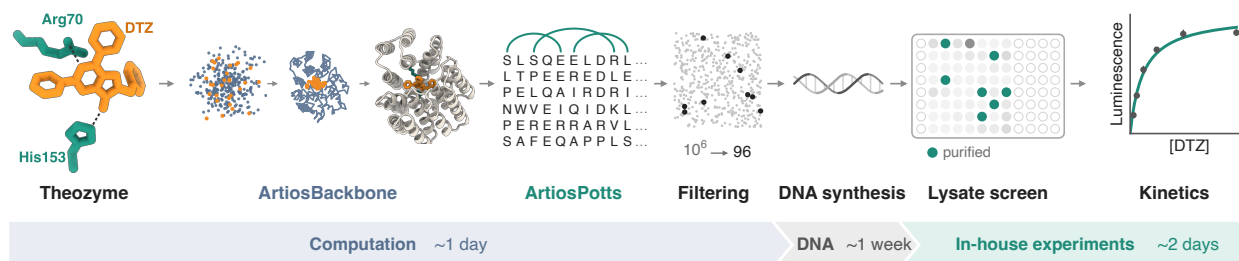


Artios: Designing Functional De Novo Enzymes with All-Atom Diffusion and Steerable Potts Models

Ligo Biosciences Team

Designing enzymes for any desired reaction is a long-standing goal of protein engineering, with applications across chemistry, medicine and industry. Directed evolution can optimise existing enzymes, but it depends on first finding a natural starting point with measurable activity. De novo design could remove this dependence by designing new enzymes directly from a description of the active site, yet it remains extremely challenging. Here we present Artios, a two-stage enzyme design system. ArtiosBackbone, an all-atom diffusion model, generates a protein backbone together with its ligand around an atom-level active-site motif and chooses where the catalytic residues sit in the sequence. ArtiosPotts, a ligand-aware Potts-model sequence designer, samples whole sequences from an explicit energy, to which terms can be added at sampling time without retraining. On the Atomic Motif Enzyme (AME) benchmark, Artios achieves state-of-the-art performance, producing up to five times as many passing designs as the strongest previous method. We also show that such terms steer ArtiosPotts away from low-complexity sequences and towards a natural amino-acid composition, and that its designed sequences are more expressible than those designed with LigandMPNN. In the laboratory, Artios produced multiple functional new-to-nature luciferases and esterases, testing no more than a single 96-well plate of designs in each campaign.



The Artios enzyme design system. A campaign starts from a theozyme (dashed lines, designed contacts). ArtiosBackbone generates the backbone together with its ligand by denoising random coordinates into a structure, and ArtiosPotts designs its sequence from an energy with pairwise couplings between positions (arcs). About 10⁶ candidates are refolded and co-folded with the substrate, and 96 designs are selected for DNA synthesis. Designs are expressed in a cell-free system and screened in lysate (shown: the ArLuc lysate screen, with the purified hits in teal), and hits are purified for Michaelis–Menten kinetics (shown: purified ArLuc1, as in Figure 4F). Computation takes about one day, DNA synthesis about one week, and in-house experiments about two days. Structures are the ArLuc1 design model; sequences are the first residues of ArLuc1–5 and ArCut1.

1 Introduction

Enzymes catalyse reactions with high rates and selectivity under mild conditions, but for most reactions of interest no suitable natural enzyme exists. Designing one from scratch means placing catalytic side chains around the substrate’s reacting atoms with sub-Ångström precision, in a pocket that holds the substrate in a reactive pose, within a protein that folds and expresses.

Deep generative models now scaffold functional motifs^{1–3}, including unindexed, atom-level active sites and all-atom complexes^{4–6}. Sequences are usually designed with ProteinMPNN or Lig-

andMPNN^{7,8}, although Potts-model sequence design methods such as Chroma and Caliby improve the consistency between designed sequences and their target structures^{9–11}. These tools have produced active de novo luciferases¹², cysteine hydrolases^{4,5} and a metallohydrolase¹³, and AlphaProtein Novo, which is not commercially available, recently designed nitrene transferases and serine esterases with catalytic efficiencies above $10^5 \text{ M}^{-1} \text{ s}^{-1}$ ¹⁴.

Three problems remain open. First, the standard enzyme motif-scaffolding benchmark scores only whether the catalytic residues are reproduced, so designs without a usable substrate pocket can pass. Second, LigandMPNN remains the standard sequence designer, and Potts-model methods had not been tested on enzymes in the laboratory. Third, finding a few active enzymes means screening hundreds of thousands of designs in silico, and telling whether a model change helps needs evaluations at a similar scale, which in turn needs efficient infrastructure.

In this research preview, we present Artios, which combines ArtiosBackbone, an all-atom diffusion model that scaffolds unindexed active-site motifs together with their ligands, and ArtiosPotts, a ligand-aware Potts-model sequence designer that can be steered at sampling time, on a modular infrastructure that scales automatically (Figure 1A–C). On a stricter motif-scaffolding benchmark that requires the designed pocket to keep its ligand, Artios passes five times as many designs as the strongest previous pipeline (Figure 1D–G). On held-out natural complexes, ArtiosPotts solves three times as many structures as LigandMPNN (Figure 2), and its sequences express more strongly in a cell-free system (Figure 3). In the laboratory, Artios produced active luciferases (7 of 85 designs tested) and cysteine esterases (17 of 88); the purified esterase ArCut1 hydrolyses 4MU-octanoate with $k_{\text{cat}}/K_M = 1.8 \times 10^3 \text{ M}^{-1} \text{ s}^{-1}$ (Figures 4 and 5). We release the Artios system and the infrastructure that runs it.

2 The Artios design system

Artios turns a description of an active site into a short list of DNA sequences to test. A campaign starts from a theozyme: a model of the reaction’s transition state or a key intermediate, together with the catalytic groups that should stabilise it. From the theozyme we keep only the atoms needed to define the catalytic geometry, typically the reacting atoms of the substrate and a few tip atoms of each catalytic side chain. ArtiosBackbone generates backbones that hold this motif, ArtiosPotts designs their sequences, and the designs are refolded with AlphaFold2¹⁵ and co-folded with the substrate using a co-folding model such as Protenix. Designs that keep the intended fold and active-site geometry are expressed in a cell-free system and screened in lysate, and hits are purified for kinetic measurements. The two models are shown in Figure 1A and B and described below, followed by the infrastructure that runs every step of a campaign (Figure 1C). Full model and training details are in the Supplementary Information.

ArtiosBackbone. ArtiosBackbone is a diffusion model that generates the protein backbone and the ligand together, without side chains outside the motif. It uses a pair-biased diffusion transformer in which backbone residues are represented by one token each and ligand atoms and motif side-chain atoms by one token per heavy atom^{3,16} (Figure 1A). The model is conditioned on the motif’s atom types and relative coordinates but not on where the motif residues sit in the sequence, so it chooses their positions while preserving the local geometry. This unindexed scaffolding⁴ matters for enzymes, where catalytic residues must be embedded in new scaffolds rather than copied at fixed positions. An optional motif-depth input controls how deeply the active site is buried. ArtiosBackbone is trained on an equal mixture of weighted PDB structures and a distillation set of predicted structures from MGnify.

ArtiosPotts. ArtiosPotts follows Chroma and Caliby^{9–11}: a ProteinMPNN-style graph neural network^{7,8} encodes the complex, and a Potts head converts its output into per-residue fields h_i and pairwise couplings J_{ij} that define an energy over sequences (Figure 1B). Unlike LigandMPNN, which builds separate protein, ligand and protein–ligand graphs, ArtiosPotts uses one graph over

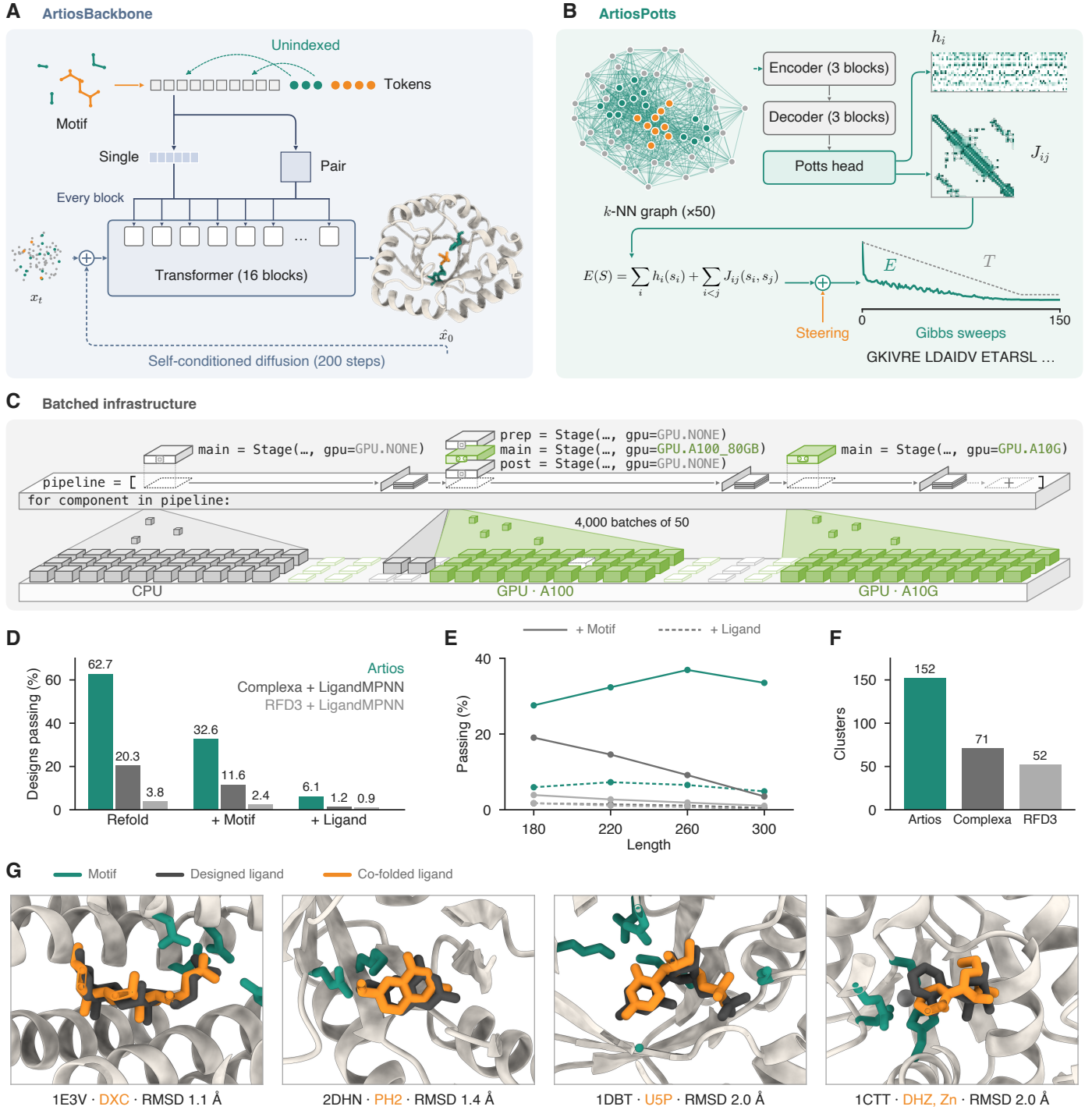


Figure 1: The Artios design system and its performance on the Atomic Motif Enzyme (AME) benchmark. **A**, ArtiosBackbone generates a backbone and its ligand around an unindexed active-site motif. **B**, ArtiosPotts predicts the fields h_i and couplings J_{ij} of a sequence energy $E(S)$ and samples sequences by annealed Gibbs sampling. **C**, The Artios infrastructure, built on Modal. A pipeline is assembled from modular components, each of which declares in code whether it runs on CPU or GPU, so every step uses only the hardware it needs; workers scale up with the number of batches and down to zero when idle. **D**, Percentage of designs passing successive AME filters (38 specifications, 200 designs each); + Ligand requires the co-folded ligand to stay in the pocket. **E**, Pass rate by design length, standard criteria (solid) and with ligand retention (dashed). **F**, Structural clusters among designs passing with ligand retention. **G**, Active sites of four passing Artios designs: co-folded ligand (orange) on the designed pose (grey) with motif side chains (teal); ligand RMSD in Å. Complexa, Proteina-Complexa; RFD3, Rfdiffusion3; both with LigandMPNN.

backbone residues, ligand heavy atoms and motif side-chain atoms. It is trained on the PDB with a composite-likelihood objective. Sequences are drawn by Markov chain Monte Carlo over the whole sequence rather than one residue at a time, using annealed, coloured Gibbs sampling as in Chroma⁹, and the fields and couplings are averaged over a small ensemble of perturbed input structures as in Caliby¹⁰.

Steering. Because ArtiosPotts samples from an explicit energy, terms can be added to that energy at sampling time without retraining (Figure 1B, steering). By default we add a Chroma-style penalty on low-complexity sequence windows⁹ and exclude cysteine and methionine at designable positions. For individual campaigns we add energy biases on single residues or residue groups, and the same mechanism supports penalties on net charge and on active-site composition. Section 3.1.3 shows how these terms control the properties of designed sequences.

Infrastructure. Every step of a campaign, from backbone generation to filtering, runs on one batched infrastructure (Figure 1C). A pipeline is written in code as an ordered list of components, and each stage of a component declares the hardware it needs, a GPU class or CPU, so moving a step between GPU and CPU, or replacing one component with another, is a one-line change. The runner executes the components in turn. For each one it sends one task per batch of designs (or per design, for some stages) to serverless workers of the declared type, which run in parallel, then waits for all batches to finish and checkpoints the results before the next component starts; a campaign can therefore be resumed or branched from any step. Workers start only when a component needs them and scale to zero when idle.

3 Results

3.1 In Silico Benchmarks

3.1.1 Enzyme motif scaffolding

We first tested whether the system can build scaffolds that hold a given active site. The Atomic Motif Enzyme (AME) benchmark contains 41 active-site specifications taken from natural enzymes, and RFdiffusion2 has been reported to scaffold all 41 of them⁴. For each one, a design pipeline must produce a protein whose sequence folds into a structure that reproduces the catalytic residues' geometry. Apart from a check that the designed structure does not clash with the ligand, the AME success criteria contain no ligand information: they test only whether the designed sequences recapitulate the motif protein residues⁴⁻⁶. While this is a necessary requirement, it neglects the ligand position after co-folding with the designed sequence. As a result, structures that are highly designable yet lack an appropriate pocket to accommodate the ligand can still be considered passing (examples in the Supplementary Information). Given that the motif scaffolding task is an assessment of a system's ability to create protein scaffolds that hold an active site, we treat such structures as failures, and accordingly propose a stricter set of filters for the AME benchmark that we argue are a more appropriate definition of success.

We append onto the standard set of AME filters a ligand positioning filter: after the designed sequence is co-folded with its ligands using Protenix, the centroid of every ligand present must lie within 3 Å of its designed position, measured after aligning on the pocket residues. This rewards structures that generate pocket environments which retain the ligand in the designed active-site geometry, ensuring that the catalytic residues and their substrate remain in the spatial arrangement necessary for enzymatic function. As in the standard criteria, a design must also refold with AlphaFold2 to within 2 Å backbone RMSD of the design with pLDDT above 80, and reproduce the motif to within 2.5 Å all-atom RMSD (Methods). We report results for 38 of the 41 specifications; three were excluded from the analysis. For each specification and pipeline, we generated 50 designs at each of four lengths (180, 220, 260 and 300 residues), giving 200 designs per specification.

We compared ArtiosBackbone with ArtiosPotts against RFdiffusion3 with LigandMPNN^{5,8} and against Proteina-Complexa with LigandMPNN⁶ (Figure 1D–G). The difference starts at the first filter: 62.7% of Artios designs refolded, against 20.3% for Proteina-Complexa and 3.8% for RFdiffusion3 (Figure 1D). Under the standard criteria, 32.6% of Artios designs passed, compared with 11.6% and 2.4%. Requiring ligand retention removed most passing designs from every pipeline, confirming that many designs reproduce the catalytic residues without a pocket that holds the substrate. Under the stricter criteria, 6.1% of Artios designs passed, five times the rate of the strongest previous pipeline (1.2% for Proteina-Complexa and 0.9% for RFdiffusion3), and Artios solved 34 of the 38 specifications, against 25 and 16.

The gap widens as designs get longer (Figure 1E). Under the standard criteria, the Artios pass rate rose from 28% at 180 residues to 37% at 260 and was 34% at 300, whereas Proteina-Complexa fell from 19% to 4% and RFdiffusion3 from 4% to 1%. With ligand retention, Artios passed 5–7% of designs at every length, while both other pipelines fell below 0.5% at 300 residues. The successful Artios designs were also more diverse: they formed 152 distinct structural clusters, against 71 and 52 (Figure 1F). Figure 1G shows the active sites of four passing Artios designs for different ligands, including a zinc-containing site (1CTT), with ligand RMSDs of 1.1 to 2.0 Å between the co-folded and designed poses.

3.1.2 Sequence design

The AME comparison changes the backbone generator and the sequence designer together. To isolate the sequence designer, we compared ArtiosPotts with LigandMPNN on fixed backbones, using natural protein–ligand complexes rather than generated structures. To avoid overlap with the training data, we built *SeqVal* from the AlphaFold3 Recent PDB evaluation set¹⁶, whose entries were released after the training cutoff and are dissimilar in sequence to the training set. After filtering for single-chain proteins with at least one buried non-covalent ligand, *SeqVal* contains 114 complexes of about 100 to 700 residues with one to four ligands (Figure 2C; Supplementary Information).

For each complex, we designed ten sequences with each model and co-folded them with their ligands. A design passed if its predicted structure matched the input backbone with high confidence and every ligand stayed in the native pocket (within 3 Å of its native centroid). A complex was counted as solved if at least one of its ten sequences passed. ArtiosPotts solved 23 of the 114 complexes and LigandMPNN 7 (Figure 2A), and ArtiosPotts produced 104 passing sequences against 18 for LigandMPNN. Every complex solved by LigandMPNN was also solved by ArtiosPotts. Ligand retention removed the most complexes for both models, but the larger difference came afterwards: of the complexes for which some sequence kept its ligands, a sequence that also refolded to the native backbone was found for 23 of 47 with ArtiosPotts and for only 7 of 31 with LigandMPNN. Figure 2G shows a two-ligand example, 7R0U, on which an ArtiosPotts sequence refolds onto the native structure (backbone RMSD 1.4 Å) and keeps both ligands in place, while the best LigandMPNN sequence misfolds. When the backbone check used AlphaFold2 instead of Protenix, ArtiosPotts still solved 6 complexes, while no LigandMPNN sequence passed (Figure 2B). The best of each complex’s ten ArtiosPotts sequences also had a higher Protenix pLDDT than the best LigandMPNN sequence for 99 of the 111 complexes with folding metrics, and a lower Protenix backbone RMSD for 94; with AlphaFold2 the counts were 101 and 89 (Figure 2D). Neither model solved any complex longer than 500 residues, but ArtiosPotts solved more complexes than LigandMPNN in every shorter length range (Figure 2H). We do not use native sequence recovery as a metric, because optimising for it does not improve, and can reduce, how well designed sequences fold.

The two models also write different kinds of sequence (Figure 2E). One in five LigandMPNN sequences contains a 30-residue window with fewer than seven effective amino acids, such as the alanine-rich stretch boxed in Figure 2E. Such windows occur in 1 of 111 native sequences and in no ArtiosPotts sequence, and the median complexity of ArtiosPotts sequences (9.4 effective amino acids in the least complex window) is close to that of native sequences (9.2; LigandMPNN, 8.1). This is partly by design, because ArtiosPotts samples with a low-complexity penalty by default (Section 3.1.3). Composition is where ArtiosPotts departs further from nature. Both models enrich

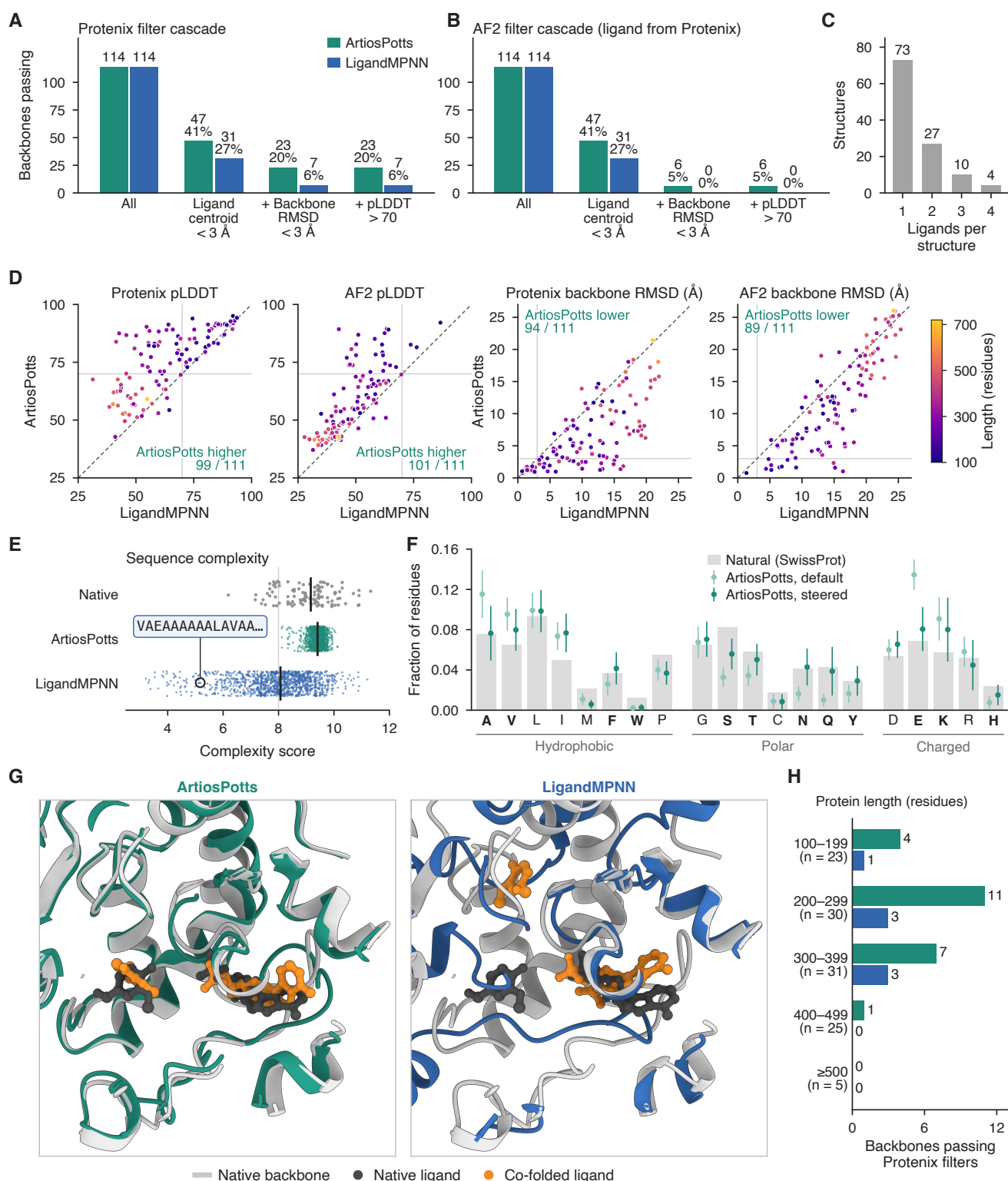


Figure 2: ArtiosPotts designs sequences for natural protein–ligand complexes that refold and keep their ligands more often than LigandMPNN. Ten sequences per model for each of the 114 *SeqVal* complexes; a complex is solved if at least one sequence passes all filters (Supplementary Information). **A**, Complexes passing successive filters, with refolding judged by Protenix. **B**, As **A**, with backbone RMSD and pLDDT from AlphaFold2, which does not see the ligand. **C**, Ligands per complex. **D**, Best of ten sequences per complex, coloured by length; counts, complexes on which ArtiosPotts is better. **E**, Least complex 30-residue window of each sequence (effective number of amino acids); lines, medians; boxed, a LigandMPNN window for 7UQ1. **F**, Composition of ArtiosPotts AME designs without and with energy biases (bold residues); grey, SwissProt. **G**, 7R0U co-folded with an ArtiosPotts sequence and the best LigandMPNN sequence (backbone RMSD 1.4 and 16.0 Å). **H**, Solved complexes by protein length.

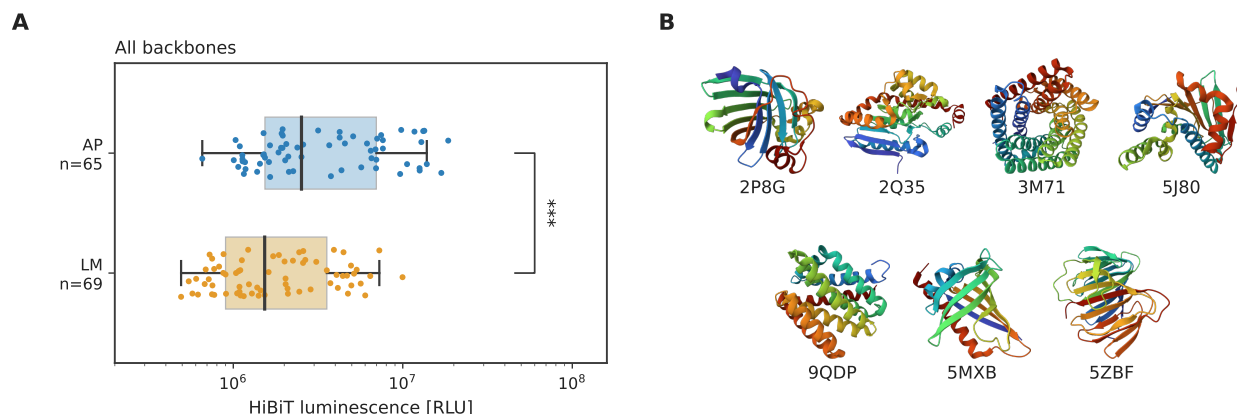


Figure 3: Cell-free expression of ArtiosPotts (AP) and LigandMPNN (LM) sequences. **A**, HiBiT luminescence of cell-free expression reactions for all sequences designed on the seven backbones in **B**, pooled. Boxes span the interquartile range with the median marked; whiskers extend to $1.5\times$ the interquartile range; points are individual sequences. The bracket shows a two-sided Mann–Whitney U test on the pooled sequences (***, $p = 6 \times 10^{-5}$). **B**, The seven natural backbones from the PDB on which both models designed sequences.

glutamate and lysine and deplete polar residues such as serine and glutamine relative to native sequences, and ArtiosPotts does so more strongly: glutamate and lysine make up 22% of its residues, against 15% for LigandMPNN and 12% for native sequences.

3.1.3 Steering sequence properties at sampling time

Properties of a designed sequence such as its complexity, composition and net charge affect whether it expresses, stays soluble and avoids aggregation. Sequence designers can have characteristic pathologies here: LigandMPNN, for example, often writes repetitive, alanine-rich stretches (Figure 2E). Because ArtiosPotts defines an explicit energy over sequences, such properties can be corrected at sampling time by adding terms to that energy, without retraining. The low-complexity penalty that ArtiosPotts uses by default is one such term: it penalises sequence windows made of only a few amino acids, and no ArtiosPotts sequence on *SeqVal* contains the low-complexity windows found in one in five LigandMPNN sequences (Figure 2E). The same mechanism supports penalties on deviation from a target composition or net charge, and energy biases on individual residues or residue groups. Without biases, ArtiosPotts over-uses glutamate, lysine and alanine (Section 3.1.2). On the 41 AME targets, biases against these residues and in favour of polar and aromatic residues brought the composition close to that of natural proteins, halving its deviation from SwissProt and bringing alanine down to its natural level, while the AlphaFold2 pass rate fell only slightly, from 83% to 79% (Figure 2F).

3.2 Experimental Validation

3.2.1 ArtiosPotts sequences show higher cell-free expression than LigandMPNN sequences

A designed enzyme is only useful if it can be produced. In silico benchmarks score whether a sequence is predicted to fold, but not whether it expresses. To compare the two sequence models directly in the laboratory, we took seven natural backbones from the PDB (2P8G, 2Q35, 3M71, 5J80, 5MXB, 5ZBF and 9QDP; Figure 3B), designed ten sequences for each with ArtiosPotts and ten with LigandMPNN, and expressed them in a cell-free system. We measured expression by HiBiT luminescence, with one well per sequence. Six of the 140 wells, marked in the plate map as having no template DNA (four) or as not available (two), were excluded, leaving 65 ArtiosPotts and 69 LigandMPNN sequences.

Pooled across all seven backbones, ArtiosPotts sequences expressed more strongly: the median

luminescence was 2.53×10^6 relative light units (RLU) for ArtiosPotts and 1.53×10^6 RLU for LigandMPNN, a 1.6-fold difference (two-sided Mann–Whitney U test, $p = 6 \times 10^{-5}$; Figure 3A).

Overall, ArtiosPotts sequences expressed better than LigandMPNN sequences designed on the same backbones.

3.2.2 Design of de novo luciferases

To validate an early prototype of our de novo enzyme design system, we first applied it to the design of a luciferase catalyzing the oxidation of diphenylterazine (DTZ) (Figure 4A). This reaction was selected as an initial benchmark because it provides a direct and highly sensitive luminescence readout and had previously been shown to be accessible to computational enzyme design¹². For this design campaign, we used the motif described by Yeh et al. and selected 85 designs produced by our model. Proteins were expressed in a cell-free system and evaluated in the lysate using a luminescence-based activity assay. Seven candidates had detectable activity, and we obtained Michaelis–Menten kinetics for the most active design after purification (Supplementary Information). The luciferase hits were diverse in both sequence and structure, and all but one adopted new-to-nature folds (Supplementary Information). The exception adopted a natural calycin/iLBP-like β -barrel fold, with comparable structural similarity to NanoLuc (TM-align TM-score ~ 0.75). The model was not given this or any other fold, so this is a convergence on a natural luciferase-like scaffold rather than a recapitulation of one.

We next asked whether Artios could design larger luciferases. In a later campaign, we screened 45 designs of 300 to 400 residues in cell-free lysate with 25 μM DTZ (Figure 4E). Nine designs were active in the initial screen, and we purified the seven strongest, which we name ArLuc1–7 in order of lysate signal. This campaign used a more developed version of the system than the first, which is why its hit rate was higher (9 of 45, compared with 7 of 85) despite the larger designs. The most active, ArLuc1, is a 300-residue helical protein (Figure 4B). In its design model, DTZ sits in a closed pocket that buries 94% of its solvent-accessible surface (Figure 4C), and the two residues specified by the motif are in place: His153 hydrogen-bonds the DTZ carbonyl oxygen (2.4 Å), and the Arg70 guanidinium sits 3.4 Å from N7, the nitrogen that is deprotonated to form the reactive anion (Figure 4D).

Purified ArLuc1 showed saturable kinetics with $K_M = 3.08 \mu\text{M}$, lower than that of LuxSit, the de novo luciferase of Yeh et al.¹², which we measured in the same experiment at the same enzyme concentration ($K_M = 7.84 \mu\text{M}$; Figure 4F). LuxSit gave more light at high DTZ concentrations, with a highest signal of 290,547 RLU against 119,770 RLU for ArLuc1, while at 0.8 and 1.6 μM DTZ the two enzymes gave similar signals. Because we recorded luminescence in relative light units without calibrating photon counts, the maximal rate cannot be converted into a turnover number, so we report K_M but not k_{cat} . LuxSit was the most active of three hits among 7,648 designs screened¹², whereas ArLuc1 came from a screen of 45 designs and is 300 residues long, against 117 for LuxSit. Yeh et al. then raised the photon flux of LuxSit more than 100-fold by site-saturation mutagenesis, giving LuxSit-i with $k_{\text{cat}}/K_M = 10^6 \text{ M}^{-1} \text{ s}^{-1}$ ¹²; ArLuc1 has not been optimised.

To test whether this activity depends on the designed active site, we mutated His153 and Arg70 of ArLuc1 to alanine, singly and together, and expressed the variants in cell-free lysate (Figure 4G). Each mutation reduced luminescence to the level of the no-enzyme blank: 14-fold for H153A, 12-fold for R70A and 30-fold for the double mutant. ArLuc1 therefore needs both catalytic residues that the motif specified.

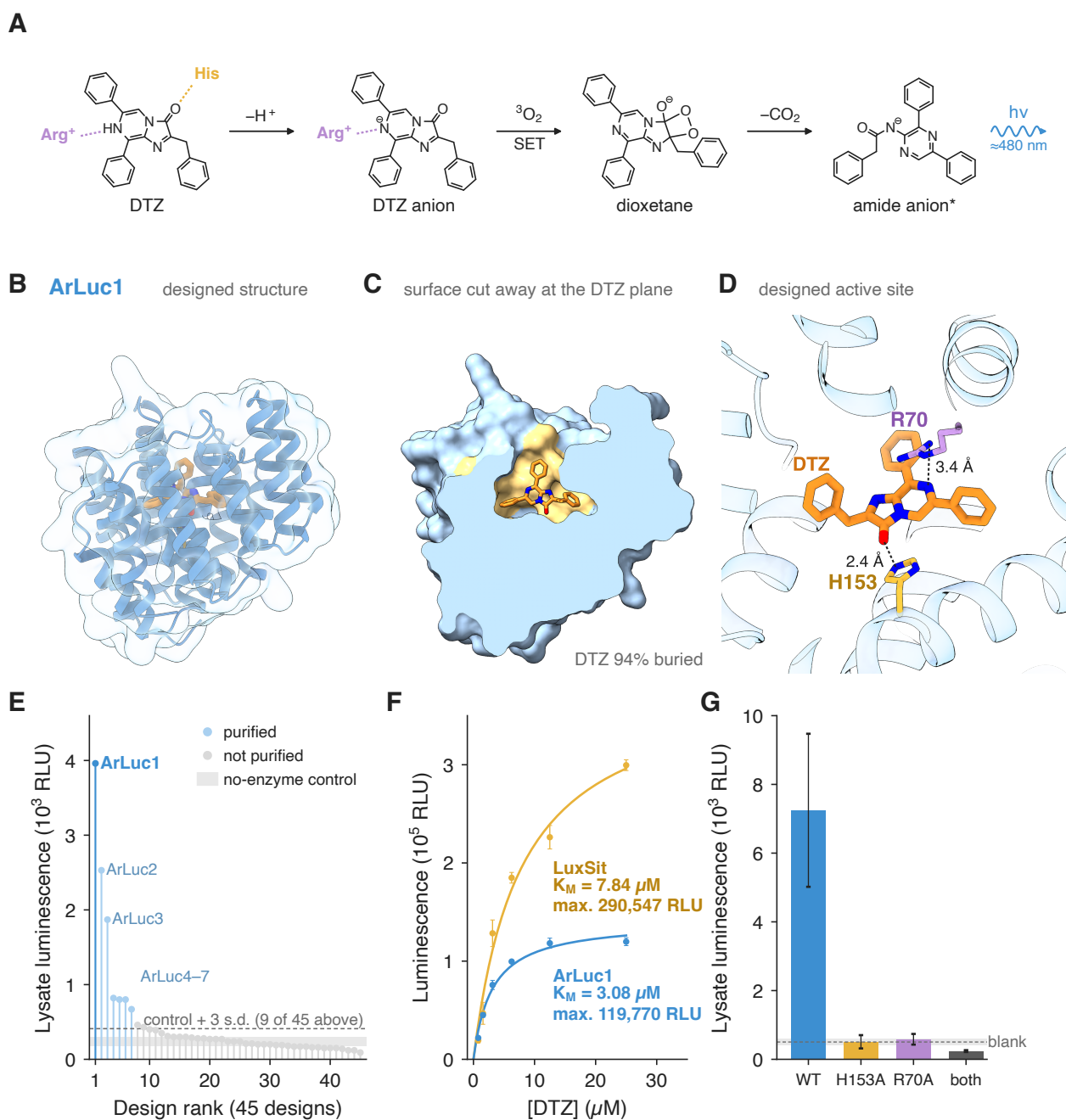
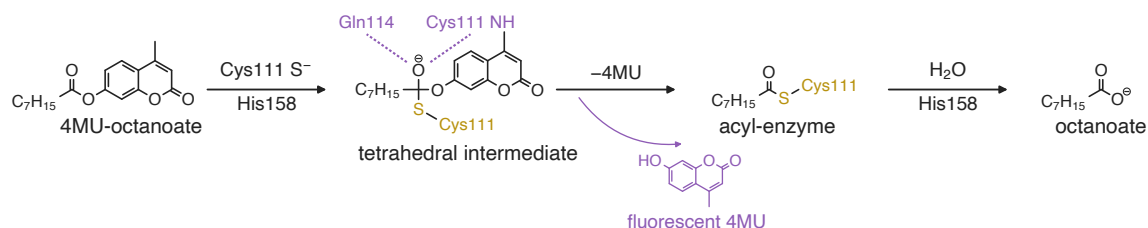


Figure 4: ArLuc1, a de novo luciferase designed by Artios. **A**, Proposed mechanism of DTZ chemiluminescence¹². The motif places an arginine at the N7 anion and a histidine at the carbonyl oxygen. The anion reacts with triplet oxygen by single-electron transfer (SET), and the resulting dioxetane loses CO₂ to give the excited amide anion, which emits at about 480 nm. **B**, Design model of ArLuc1 predicted with Protenix: cartoon inside a transparent molecular envelope (6 Å resolution), with DTZ in orange and the residues lining its pocket tinted cream. **C**, Molecular surface cut away just in front of DTZ, showing the substrate in its pocket (cream, residues within 4.5 Å). 94% of the solvent-accessible surface of DTZ is buried. **D**, Designed active site, with the catalytic His153 (gold) and Arg70 (lilac); distances in Å. **E**, Cell-free lysate screen of 45 designs with 25 μM DTZ, one well per design, ranked by luminescence. Grey band, mean ± s.d. of seven no-enzyme controls; dashed line, mean + 3 s.d. ArLuc1–7 are the seven designs that were purified. **F**, Luminescence of 200 nM purified ArLuc1 and LuxSit¹² at 0.8–25 μM DTZ, minus the buffer blank (mean ± s.d., three wells), with Michaelis–Menten fits. **G**, Alanine variants of the catalytic residues in cell-free lysate (25 μM DTZ; mean ± s.d. of three reads of one well). Dashed line and grey band, blank wells. WT, wild type.

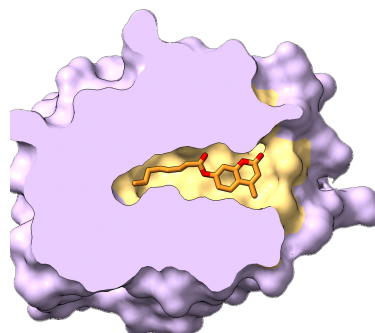
A



B ArCut1 designed structure

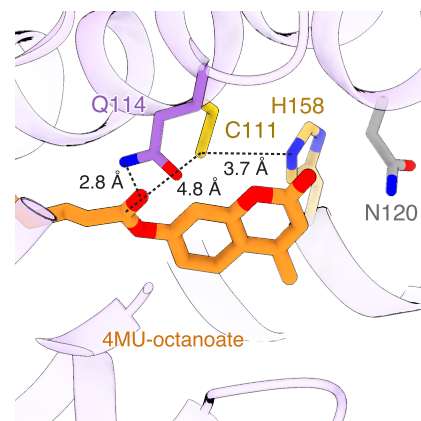


C surface cut at the substrate plane

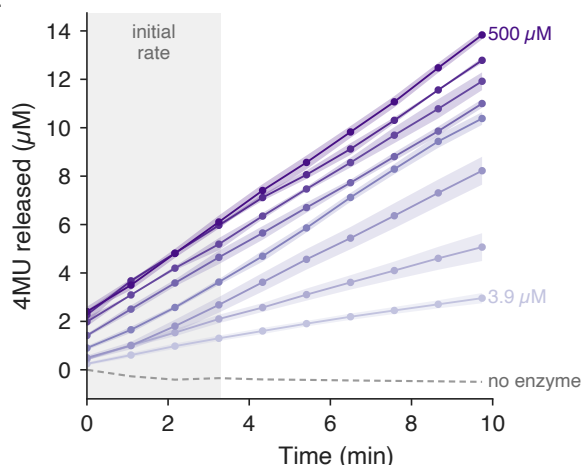


substrate 96% buried

D designed active site



E



F

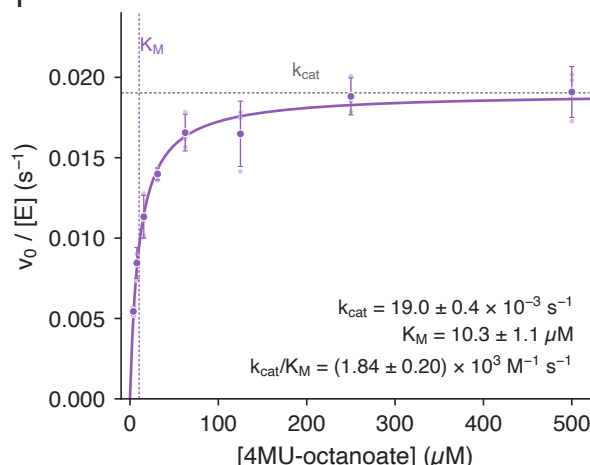


Figure 5: ArCut1, a de novo cysteine hydrolase for 4MU-octanoate. **A**, Proposed mechanism. His158 activates Cys111, whose thiolate attacks the ester carbonyl; the oxyanion of the tetrahedral intermediate is stabilised by Gln114 and the Cys111 backbone amide. Collapse of the intermediate releases fluorescent 4-methylumbelliferone (4MU) and leaves an acyl-enzyme, which water hydrolyses to octanoate. C₇H₁₅, heptyl chain. **B**, Design model of ArCut1 predicted with Protenix: cartoon inside a transparent molecular envelope (6 Å resolution), with 4MU-octanoate in orange and its pocket lining tinted cream. **C**, Molecular surface cut away just in front of the substrate. The octanoyl chain lies in a deep, buried channel (cream, residues within 4.5 Å); 96% of the substrate's solvent-accessible surface is buried. **D**, Designed active site: Cys111, His158, Gln114 and Asn120; distances in Å. **E**, 4MU released by 1 μM purified ArCut1 at eight 4MU-octanoate concentrations (3.9–500 μM, two-fold steps), after subtraction of substrate-only wells (mean ± s.d., three wells). Shaded, the first 3.3 min used for initial rates. Grey dashed line, substrate-only wells at 500 μM. **F**, Initial rate per enzyme against substrate concentration, with a Michaelis–Menten fit (mean ± s.d.; light points, individual wells). Dotted lines mark k_{cat} and K_M .

3.2.3 Design of de novo short- and medium-chain esterases

For our next design campaign, we selected ester hydrolysis because esterase catalysis is mechanistically well-characterized and these enzymes perform many cellular functions and have numerous industrial applications. To design the minimal catalytic components for this reaction, we extracted the catalytic residues and positions from a crystal structure of a Papaya cysteine hydrolase (PDB: 1PPN)¹⁷. We placed atoms from each of the cysteine, histidine, and asparagine residues to specify the relative geometry of the catalytic triad, together with one atom of a glutamine positioned to form the oxyanion hole. We then oriented a conformer of our reactant (4MU-butyrate) according to the known optimal geometries for nucleophilic attack by the cysteine, proton donation by the histidine, and hydrogen bond stabilization of the oxyanion^{4,18}. Among the 88 designs that were experimentally screened, 17 showed detectable activity (Supplementary Information); the best design exhibited multiple-turnover activity with a k_{cat}/K_M of $291 \text{ M}^{-1} \text{ s}^{-1}$ for the acylation step. RFdiffusion2, starting from the same papain structure and substrate, reported a best design with $248 \text{ M}^{-1} \text{ s}^{-1}$ from 48 designs screened⁴.

Having established esterase activity on 4MU-butyrate, we next extended the approach to a longer-chain ester, 4-methylumbelliferyl octanoate (4MU-octanoate). This is a more stringent test of shape complementarity, because the pocket must also hold the eight-carbon acyl chain. Hydrolysis follows the cysteine-hydrolase mechanism: His158 activates Cys111, whose thiolate attacks the ester carbonyl, and collapse of the tetrahedral intermediate releases fluorescent 4-methylumbelliferone (4MU) (Figure 5A,D). We purified the most active design, ArCut1, a 170-residue helical protein with $k_{\text{cat}}/K_M = (1.84 \pm 0.20) \times 10^3 \text{ M}^{-1} \text{ s}^{-1}$ (Figure 5B,E,F). In its design model, the octanoyl chain lies in a buried hydrophobic channel, and 96% of the substrate’s solvent-accessible surface is buried (Figure 5C). This activity is of the same order as the best designed hydrolases that did not go through experimental optimisation. RFdiffusion3 found 35 multiple-turnover cysteine hydrolases among 190 designs, the best with k_{cat}/K_M of about $3.6 \times 10^3 \text{ M}^{-1} \text{ s}^{-1}$ on the activated substrate 4MU-phenylacetate⁵. Designed serine hydrolases reached $3.8 \times 10^3 \text{ M}^{-1} \text{ s}^{-1}$ directly from design, and $2.2 \times 10^5 \text{ M}^{-1} \text{ s}^{-1}$ after a further round of redesign¹⁹.

3.2.4 Summary across campaigns

Table 1 summarises the scale and outcome of each campaign.

Table 1: Summary of experimental campaigns.

Campaign	Backbones	Screened	Hits
Luciferase (DTZ)	10^6	85	7
Luciferase, large (DTZ)	10^6	45	9
Esterase (4MU-butyrate)	10^6	88	17
Esterase (4MU-octanoate)	10^6	88	1

4 Discussion

Motif-scaffolding benchmarks have guided much of the recent progress in generative protein design, yet the standard enzyme benchmark evaluates only whether the catalytic residues are reproduced. When we additionally required the designed pocket to retain its ligand after co-folding, most of the designs that satisfied the standard criteria failed, for every pipeline we tested. Recapitulating an active-site geometry and building a scaffold that can hold the substrate are therefore distinct problems, and we suggest that the stricter criterion is a more informative target for the development of enzyme design methods, irrespective of the models presented here. Most de novo designs are far below the size of natural enzymes, because sequence design models struggle at longer lengths; Artios maintained its pass rate up to the longest designs we benchmarked and produced active luciferases of up to 400 residues, bringing de novo enzymes closer to the size of their natural counterparts.

To our knowledge, this work also provides the first laboratory evaluation of a Potts-model sequence designer on enzymes, in which ArtiosPotts outperformed LigandMPNN both in silico and in cell-free expression. An equally important property of the approach is that an explicit sequence energy makes the sequence designer controllable. Properties that arise as side effects in other designers, such as low-complexity stretches or a skewed amino-acid composition, can instead be specified as terms in the energy at sampling time without retraining, so that the sequence-level requirements of a campaign enter the design process directly rather than being imposed by filtering afterwards. The comparisons reported here, which required thousands of designs per pipeline, were made practical by an infrastructure in which each step of a campaign runs in batches on the hardware it requires and scales with demand. This allowed changes to the models to be assessed on samples large enough to be informative, and we release the infrastructure together with the models.

The principal limitation of this work lies in the selection of designs. Our filters depend on structure predictors, namely AlphaFold2 and Protenix, which remain crude relative to the requirements of catalysis. Selection for confident refolding favours simpler folds and therefore biases the set of designs tested towards them, while co-folding indicates whether a ligand occupies the designed pocket but does not resolve the precise geometry and interactions on which a catalytic mechanism depends. The measured activities remain well below those of natural enzymes and of designed enzymes that have been optimised in the laboratory, and experimental structures of the designs are not yet available. We will continue to validate the system on new reactions.

5 Contributors

Authors are listed in alphabetical order of surname: Oliver Cort, Emily Egerton-Warburton, Arda Goreci, Edward Harris.

A.G. and O.C. built the computational system and trained the models. E.E.-W. designed the active-site motifs for the enzyme campaigns and led the experimental work. E.H. performed experiments and managed the project. A.G., O.C., E.E.-W. and E.H. contributed to the manuscript.

6 Acknowledgements

We thank Maximilian Gantz (Max Planck Institute for Terrestrial Microbiology and Institute for Protein Design, University of Washington), our long-standing advisor, for guidance throughout this project. We thank Carlos Vonessen, who during an internship at Ligo Biosciences in summer 2025 wrote code for an early version of our sequence design model, which later developed into ArtiosPotts. We are grateful to Modal and to Erik Bernhardsson, who arranged our access to thousands of GPUs; this let us scale up our experiments and finish the project in a reasonable time.

7 Methods

7.1 Protein design pipeline

For each campaign, ArtiosBackbone generated up to 1,000,000 backbone–ligand complexes, split equally across three or four target lengths between 160 and 300 residues. We used 200 diffusion steps with self-conditioning and no guidance. ArtiosPotts designed one or more sequences per backbone. Sequences were folded with AlphaFold2¹⁵ and filtered on backbone RMSD to the design, pLDDT and motif geometry; designs that passed were co-folded with the substrate using Protenix and filtered on active-site geometry. In later campaigns, selected backbones were redesigned with ten new sequences each and re-evaluated. Campaign-specific settings and filter thresholds are given in the Supplementary Information.

7.2 Active-site specification

Each motif was built from a theozyme or from a natural active site. It contained the substrate (or a transition-state or intermediate analogue) and a small number of atoms from each catalytic residue, usually the side-chain tip atoms, placed in the geometry required for catalysis. Sequence positions of the motif residues were not specified. Motifs for each campaign are described in the Supplementary Information.

7.3 Sequence design

ArtiosPotts sequences were sampled by coloured Gibbs sampling with the temperature annealed linearly from 1.0 to 0.1 over the first 80% of sweeps, using fields and couplings averaged over an ensemble of noise-perturbed input structures. Unless stated otherwise, we applied a low-complexity penalty and excluded cysteine and methionine at designable positions, while motif residues were kept fixed.

7.4 Structure prediction and filtering

Designed sequences were folded with AlphaFold2 without the ligand and co-folded with their ligands using Protenix. For the AME benchmark, a design passed if (i) the AlphaFold2 structure was within 2.0 Å backbone RMSD of the design with pLDDT above 80, (ii) the motif all-atom RMSD of the AlphaFold2 structure was below 2.5 Å, and (iii) after aligning on pocket residues, every ligand centroid in the Protenix co-folded structure was within 3.0 Å of its designed position. Pocket residues were those with a heavy atom within 8 Å of any ligand atom.

7.5 DNA synthesis and amplification

Coding sequences were codon-optimised for *Escherichia coli* with Benchling and ordered as linear DNA from Abstract Biosciences or Twist Bioscience. Constructs contained a T7 promoter, a C-terminal 6×His tag and a T7 terminator. Genes were amplified with Q5 High-Fidelity DNA Polymerase (NEB) and purified with the Monarch Mag PCR and DNA Cleanup Kit (NEB).

7.6 Cell-free protein expression

Linear templates were expressed with the CFXpress cell-free expression system (GenScript) supplemented with GamS (NEB), according to the manufacturer’s instructions. For the expression study, 5 µL of cell-free reaction diluted tenfold was mixed with 5 µL of HiBiT detection reagent in a 384-well plate, and luminescence was read on a GloMax plate reader (Promega).

7.7 Protein purification

Proteins were purified by Ni-NTA affinity chromatography in 96-well format following the Semi-Automated Protein Production workflow²⁰, buffer-exchanged into PBS using AcroPrep Advance 96-well filter plates (Cytiva), and checked by SDS-PAGE.

7.8 Enzyme activity assays

Luciferase. Lysate screens followed Yeh et al.¹²: 12.5 µL of cell-free lysate was mixed with 25 µM DTZ in a 50 µL reaction in white half-area 96-well plates, and endpoint luminescence was read on a GloMax plate reader. For kinetics, purified enzyme at a final concentration of 1 µM was mixed with 0.39–50 µM DTZ, luminescence was read every minute for 10 minutes, and initial rates from the first three reads were fitted to the Michaelis–Menten equation. For ArLuc1–7 (Figure 4), purified enzyme was used at 200 nM final concentration with DTZ from 0.78 to 100 µM in two-fold steps, in triplicate wells. The buffer blank for each DTZ concentration was subtracted, and Michaelis–Menten

fits used concentrations up to the one giving the highest signal. Alanine variants were expressed in cell-free lysate and read three times, 35 s apart, after adding 25 μ M DTZ.

Esterase. Lysate assays used 10 μ L of cell-free lysate and purified assays used 1 μ M final enzyme concentration in 100 μ L of PBS containing 5% DMSO, in black 96-well plates. Lysate screens used 100 μ M 4MU-butyrate or 4MU-octanoate, and kinetic assays used 3.91–1000 μ M 4MU-butyrate or 1.95–500 μ M 4MU-octanoate. Fluorescence (excitation 365 nm, emission 415–445 nm) was read every minute for 60 minutes. Signals were corrected using substrate-only wells and compared with no-DNA cell-free controls.

7.9 Data processing and statistical analysis

In lysate screens, a design was called a hit if its signal exceeded the mean of the matched negative controls by more than three standard deviations. For the luciferase screens, which used replicate wells, we also required the replicate mean minus one standard deviation to exceed this threshold. For the expression study, ArtiosPotts and LigandMPNN luminescence values were compared with a two-sided Mann–Whitney U test on the sequences pooled across the seven backbones, using the last of three plate reads, taken 16 minutes after the first. The two earlier reads show differences in the same direction.

8 Code availability

The Artios models and design infrastructure are released as open source alongside this paper.

9 Supplementary Information

Supplementary Information (data pipeline, model details, benchmarks, full design campaigns, experimental methods and reference tables) is provided as a separate document.

References

- [1] Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Nikita Hanikel, Samuel J. Pellock, Alexis Courbet, William Sheffler, Jue Wang, Preetham Venkatesh, Isaac Sappington, Susana Vázquez Torres, Anna Lauko, Valentin De Bortoli, Emile Mathieu, Sergey Ovchinnikov, Regina Barzilay, Tommi S. Jaakkola, Frank DiMaio, Minkyung Baek, and David Baker. De novo design of protein structure and function with RFdiffusion. *Nature*, 620(7976):1089–1100, 2023. doi: 10.1038/s41586-023-06415-8.
- [2] Yeqing Lin, Minji Lee, Zhao Zhang, and Mohammed AlQuraishi. Out of many, one: Designing and scaffolding proteins at the scale of the structural universe with Genie 2. *arXiv preprint arXiv:2405.15489*, 2024. doi: 10.48550/arXiv.2405.15489.
- [3] Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Proteina: Scaling flow-based protein structure generative models. *arXiv preprint arXiv:2503.00710*, 2025. doi: 10.48550/arXiv.2503.00710. ICLR 2025.
- [4] Woody Ahern, Jason Yim, Doug Tischer, Saman Salike, Seth M Woodbury, Donghyo Kim, Indrek Kalvet, Yakov Kipnis, Brian Coventry, Han Raut Altae-Tran, et al. Atom-level enzyme active site scaffolding using RFdiffusion2. *Nature Methods*, 23:96–105, 2026. doi: 10.1038/s41592-025-02975-x.

- [5] Jasper Butcher, Rohith Krishna, Raktim Mitra, Rafael I. Brent, Yanjing Li, Nathaniel Corley, Paul Kim, Jonathan Funk, Simon Mathis, Saman Salike, Aiko Muraishi, Helen Eisenach, Tuscan Rock Thompson, Jie Chen, Yuliya Politanska, Enisha Sehgal, Brian Coventry, Odin Zhang, Bo Qiang, Kieran Didi, Max Kazman, Frank DiMaio, and David Baker. De novo design of all-atom biomolecular interactions with RFDiffusion3. *bioRxiv*, 2025. doi: 10.1101/2025.09.18.676967.
- [6] Kieran Didi, Zuobai Zhang, Guoqing Zhou, Danny Reidenbach, Zhonglin Cao, Sooyoung Cha, Tomas Geffner, Christian Dallago, Jian Tang, Michael M. Bronstein, Martin Steinegger, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Scaling atomistic protein binder design with generative pretraining and test-time compute. *arXiv preprint arXiv:2603.27950*, 2026. doi: 10.48550/arXiv.2603.27950.
- [7] Justas Dauparas, Ivan Anishchenko, Nathaniel Bennett, Hua Bai, Robert J. Ragotte, Lukas F. Milles, Basile I. M. Wicky, Alexis Courbet, Rob J. de Haas, Neville Bethel, Philip J. Y. Leung, Timothy F. Huddy, Samuel Pellock, Doug Tischer, Frederick Chan, Brian Koepnick, Hannah Nguyen, Alex Kang, B. Sankaran, Asim K. Bera, Neil P. King, and David Baker. Robust deep learning-based protein sequence design using ProteinMPNN. *Science*, 378(6615):49–56, 2022. doi: 10.1126/science.add2187.
- [8] Justas Dauparas, Gyu Rie Lee, Robert Pecoraro, Linna An, Ivan Anishchenko, Cameron Glasscock, and David Baker. Atomic context-conditioned protein sequence design using LigandMPNN. *Nature Methods*, 22:717–723, 2025. doi: 10.1038/s41592-025-02626-1.
- [9] John B. Ingraham, Max Baranov, Zachary Costello, et al. Illuminating protein space with a programmable generative model. *Nature*, 623:1070–1078, 2023. doi: 10.1038/s41586-023-06728-8.
- [10] Richard W. Shuai, Tianyu Lu, Subhang Bhatti, Petr Kouba, and Po-Ssu Huang. Ensemble-conditioned protein sequence design with Caliby. *bioRxiv*, 2025. doi: 10.1101/2025.09.30.679633. Preprint.
- [11] Jinho Kim, Richard W. Shuai, and Po-Ssu Huang. Two-stage ligand-protein complex sequence design with L-Caliby. In *ICLR 2026 Workshop on Foundation Models for Science*, 2026. URL <https://openreview.net/forum?id=NcQZyVzyMf>.
- [12] Andy Hsien-Wei Yeh, Christoffer Norn, Yakov Kipnis, Doug Tischer, Samuel J Pellock, Declan Evans, Pengchen Ma, Gyu Rie Lee, Jason Z Zhang, Ivan Anishchenko, et al. De novo design of luciferases using deep learning. *Nature*, 614(7949):774–780, 2023. doi: 10.1038/s41586-023-05696-3.
- [13] Gina El Nesr, Simon L. Dürr, Irimpan I. Mathews, Qi Wen, Kewei Zhao, Ritimukta Sarangi, Ursula Röthlisberger, Fanny Sunden, and Po-Ssu Huang. Zero-shot design of a de novo metalloenzyme. *bioRxiv*, 2026. doi: 10.64898/2026.04.23.720277.
- [14] Zachary Wu, Josh Abramson, Thomas Frerix, Alexander E. Chu, Ruijie K. Zhang, et al. Designing enzymes for new-to-nature chemistry and non-natural substrates with AlphaProtein Novo. *bioRxiv*, 2026. doi: 10.64898/2026.10.01.756017.
- [15] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Židek, Anna Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.
- [16] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630:493–500, 2024. doi: 10.1038/s41586-024-07487-w.

- [17] R. W. Pickersgill, G. W. Harris, and E. Garman. Structure of monoclinic papain at 1.60 Å resolution. *Acta Crystallographica Section B*, 48:59–67, 1992.
- [18] Andrew R Buller and Craig A Townsend. Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad. *Proceedings of the National Academy of Sciences*, 110(8):E653–E661, 2013. doi: 10.1073/pnas.1221050110.
- [19] Anna Lauko, Samuel J Pellock, Kiera H Sumida, Ivan Anishchenko, David Juergens, Woody Ahern, Jihun Jeung, Alexander F Shida, Andrew Hunt, Indrek Kalvet, Christoffer Norn, Ian R Humphreys, Cooper Jamieson, Rohith Krishna, Yakov Kipnis, Alex Kang, Evans Brackenbrough, Asim K Bera, Banumathi Sankaran, K N Houk, and David Baker. Computational design of serine hydrolases. *Science*, 388(6744):eadu2454, 2025. doi: 10.1126/science.adu2454.
- [20] Jason Qian, Lukas F. Milles, Basile I. M. Wicky, Amir Motmaen, Xinting Li, Ryan D. Kibler, Lance Stewart, and David Baker. Accelerating protein design by scaling experimental characterization. *bioRxiv*, 2025. doi: 10.1101/2025.08.05.668824.